12-3-2010

DataONE Strucure and Potential Partnership as a Member Node

DataONE

DataONE Structure and Potential Partnership as a Member Node

DataONE, a new initiative funded by the National Science Foundation in the United States of America aims to provide better access to biological, ecological and environmental data throughout the world. There are four challenges that DataONE is established to address:

- **data loss** – preserving the work that has been done
- **data dispersion** – facilitating discovery of data
- **data deluge** – assist stakeholders (environmental scientists, teachers, policy makers etc) navigate the flood of increasingly heterogeneous data
- **poor practice** – to provide best-practice models and training for data management and storage.

DataONE aims to do this within the USA, and in partnership with groups throughout the world to provide a global network of information about global biodiversity. Already many datasets are stored in other countries than those of their origin and this will be a way to ensure those data are discoverable by all. An exemplar for this is in some of the marine and climatic data-sharing models, where it is well-recognized that national boundaries are irrelevant or at least limiting to the better knowledge, understanding and management of our natural systems.

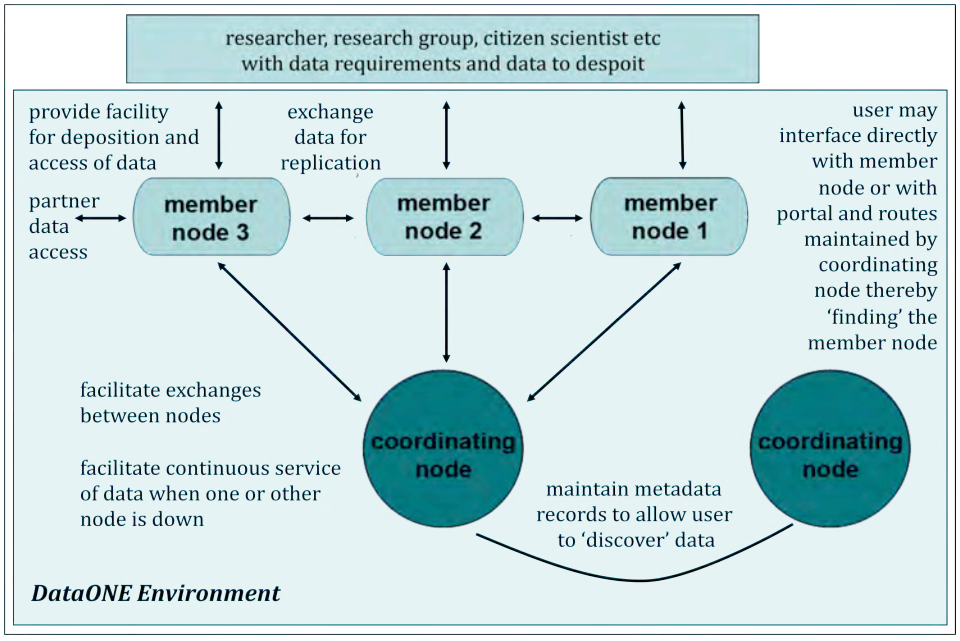


Figure 1: Depiction of the interface between client (stakeholder) and the DataONE environment. Member Nodes contain and sometimes (not always) receive data (a Coordinating Node may also have Member Node status), while Coordinating Nodes contain and transmit the knowledge of each Member Node's holdings and ensure back-up between Member Nodes,

There are three components in DataONE:

Member Nodes: located at institutions distributed throughout academia, libraries, government agencies, and other organizations that provide local data storage, curation, and metadata for a set of data resources that are collected or affiliated with that organization.

Coordinating Nodes: are geographically-distributed to provide a high-availability, fault-tolerant, and scalable set of coordinating services to the Member Nodes, including a complete metadata index and data replication services.
Investigator Toolkit: a desktop and web-based suite of software tools for researchers and the principle form of interaction with the network.

The user will interface with the data storage facilities in DataONE, either through direct pathways to individual storage institutions (the Member Nodes) or through an enquiry at the DataONE portal, which will direct the user to the appropriate location in the network (Figure 1). The value of the system is largely in the anticipated continuous service provided (no down-time due to one or other Node being off-line), and in the collegiate support provided to the network members.

The interface of DataONE with the ecosystem science community will reflect the diverse needs of stakeholders in:

1. the kind of collection (dictating the type of data held and its mode of access), ranging from species collections to data outputs from temporal data collection stations, or in
2. the manner of access. Many scientists, for example, will wish to have access to raw data, while decision-makers, such as on-ground conservation officers, need prioritized and manipulated data, in order to obtain relevant geographic information about rare species and priority for conservation. DataONE aims to provide the interface between data sources (Member Nodes) and the wide variety of users, in collaboration with other data providers such as librarians (Figure 2).

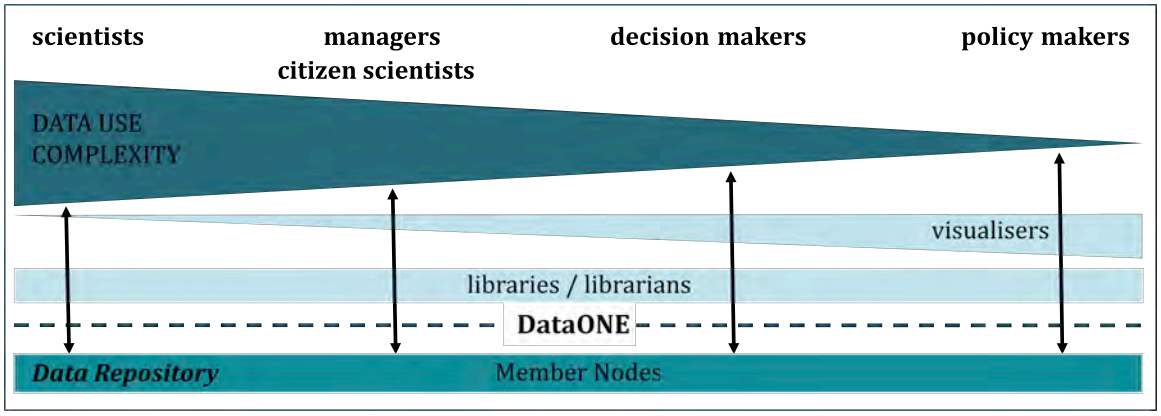


Figure 2: The conversion requirements for data according to type of audience, and the importance of data-transformation.

For data to be appropriately stored and discoverable, a good system of data description needs to be imposed. For this, metadata standards need to be established, but for the wide range of data types expected, there will be no single metadata standard and neither can one be imposed. DataONE's technical goal is to provide the background for networks to interoperate with one another with the aim of providing data valuable for researchers and other users in the ecosystem science fields. DataONE is focused on long-term preservation of data (100 years plus): how can you preserve data when institutions, governments and countries may not be around in 100 years time. Focusing on institutional diversity can be a main way to provide preservation. The aim is to provide an architecture that can pick interfaces that make sense to allow general agreement.

Much work is required on the creation and update of systems as well as extracting data: the different service interfaces required present some real challenges. The operating environment for DataONE in the first phase has three Coordinating Nodes (University of New Mexico, University of California, Santa

Barbara and Oak Ridge Research Consortium), three Member Nodes (KNB, ORNL DAAC, Dryad) and an Investigator Toolkit (v1.0).

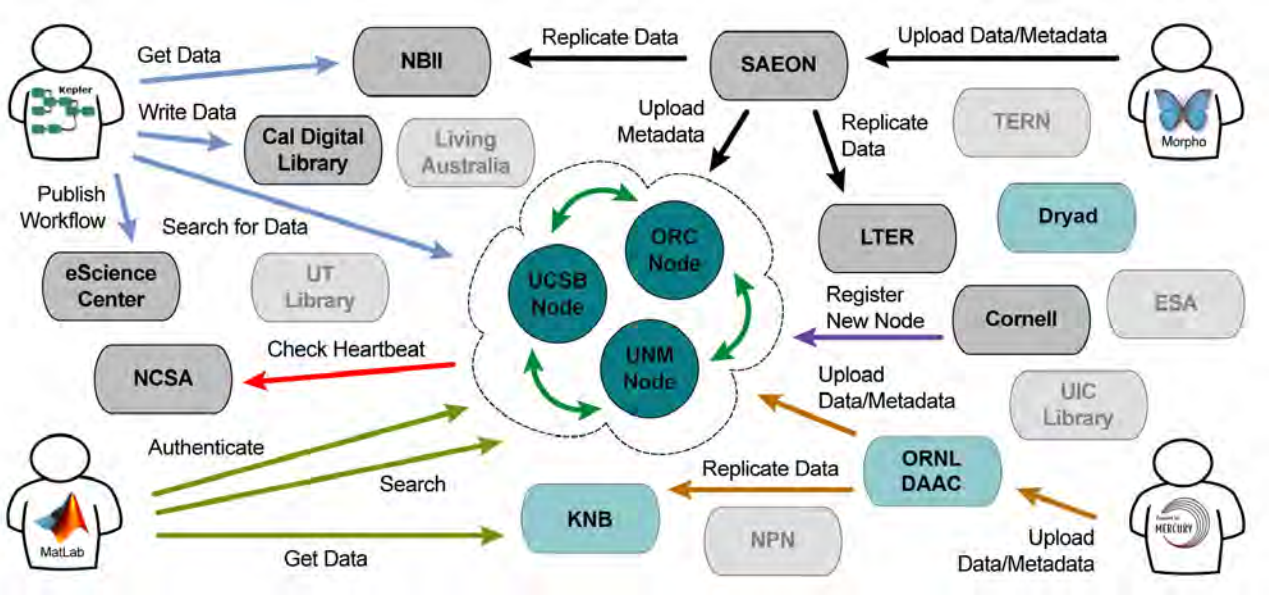


Figure 3: DataONE Member Nodes form a robust, distributed network via coordinating services provided by a set of Coordinating Nodes (central circles i.e., Oak Ridge Campus, UC Santa Barbara, and University of New Mexico) arranged in a high-availability configuration. Scientists and citizens interact with Member Nodes (e.g., South African Environmental Observation Network, California Digital Library, USGS National Biological Information Infrastructure) through software tools that utilize standardized interfaces (active Member Nodes are depicted in color, potential Member Nodes are grey). This structure supports many different usage scenarios, such as data and metadata management and replication (e.g., using Morpho [black arrows] or the Mercury system [orange arrows]), as well as analysis and modeling (e.g., using commercial software like Matlab [light green arrows] and open-source scientific workflow systems like Kepler [blue arrows]). Coordinating Nodes perform many basic indexing and data replication services to ensure data availability and preservation (e.g., node registration [purple arrow] and monitoring via heartbeat services [red arrow])

DataONE will develop a comprehensive set of tools for those wishing to access DataONE facilities, called the Investigator Toolkit. At present most scientists do not go to the web to get data (with the exception of some data types, such as meteorological information), but use their personal contacts. These bring with them personal knowledge of the data standards in the dataset imposed, but are limited in terms of range and exploration of the actual available data. There will need to be a change in practice to enable the approach proposed by DataONE to work. The vision is that this will be global.

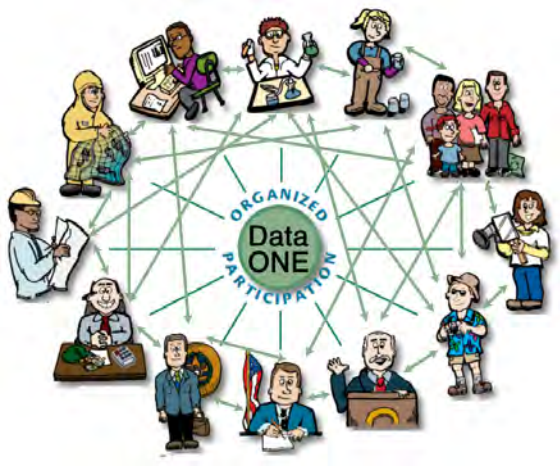
The quality of data is an important part of its value, and knowing the pathways by which the data were created is critical to this. Reproducible synthesis is critical and DataONE will work to improve practice in this regard through its training component and incorporation of workflow models such as Kepler. The crux of the approach is to try to capture the process that the scientist goes through and record it, the steps to produce any end product: the scientific workflow. There is huge heterogeneity in modeling the scientific workflow due to operating system dependencies, different modeling and programming languages, and so on. The Kepler program has been designed to assist this process, decomposing tasks within each major step. Building this can simplify the whole process.

Participating in DataONE

So You Wish to Become an Active Part of DataONE?

One of the primary ways to interact with DataONE as a partner is to become a Member Node.

Member Nodes store and provide access to their digital scientific data holdings. Member Nodes provide metadata describing their data to Coordinating Nodes to facilitate discovery and retrieval of the data within the DataONE Network. Member Nodes may host their own data, and/or may partner with other data holders to provide access to partner's data and metadata. Member Nodes may have direct connection with user communities.



Member Nodes are distinct from **Coordinating Nodes,** which choreograph services that help users store, discover, and retrieve data using metadata provided by Member Nodes, who archive data using various standards and formats. The Coordinating Nodes facilitate replication and preservation of data objects, authorization and authentication, and data discovery. They enable replication of archived data object at other geographically and organizationally distinct Member Nodes and are responsible for authentication and authorization for any operations by users or services on DataONE.

Potential Benefits for Member Nodes

- Improved methods of access to your data, as well as wider access to and exposure of those data
- Cost-effective preservation of your data and metadata
- Access to larger community with expertise and best practices in data life cycle management
- A toolkit for analysis, visualization and modeling your data
- Facilitate response to increasing demands by funding agencies for long-term data management planning
- Visibility as a community leader in supporting open and rapid access to scientific data
- Recognition and credit for data and services provided through data citations in published literature
- Opportunities for collaborative research
- Potential funding opportunities resulting from data discovery

Responsibilities

- Provide access to actively managed data
- Allow replication of data for preservation and increased access
- Provide descriptive metadata to Coordinating Nodes to facilitate data discovery, access and usability
- Ensure availability and reliability of physical data access
- Engage with the larger DataONE community to advance DataONE services, support emerging Member Nodes, and promote best practices in data management.
- Utilize a unique identifier for each dataset by participating in DataONE user identity federation
- Deploy an implementation of DataONE APIs (Application Programming Interfaces)

Criteria for Standing Up a DataONE Member Node

This document provides guidance for the DataONE project on prioritizing deployment of new Member Nodes that require some resources from the DataONE project in order become operational. It is not meant to restrict the pool of potential candidates that may bring their own resources to bear on the process of participation as a Member Node.

Size and visibility of the community represented by the candidate Member Node

The collections are significant or enable new science or both

- Fills significant gaps in the content available through DataONE
- The data are unique data in the broader community
- Collections are strong in breadth, depth, or both

The candidate brings significant contributions to the DataONE resource base, including:

- Strategic partnerships
- Professional expertise in managing data, computing, administration, etc.
- Synergistic services (e.g., TeraGrid)
- New funding streams or other sustainability enhancing resources
- Technical resources such as storage capacity, bandwidth, and processing power
- Compatibility with DataONE services, minimizing cost of deployment

The candidate adds diversity to the collective membership of DataONE such as through:

- Geographic diversity: a new state or region, new country, new continent
- Under-represented group
- Linguistic and cultural diversity
- Different type of institution
- Diversity of funding sources

The candidate offers compliance with best practices and standards, including:

- Quality assurance
- Data sharing policies
- Metadata creation
- Security

Process to Establish a Member Node

Procedure

Step 1:

Expression of interest

- Complete Member Node Description document and sign letter of interest
- Invite/Join DataONE Users Group (at any time)

Step 2:

Technical and strategic review in conjunction with Core CI Team and Leadership Team

- See 'Criteria for Standing Up a DataONE Member Node' for general information

  The intent is that all potential participants be welcomed.

Step 3:

DataONE response:

Response 1: Yes, become a Member Node

- Agree to DataONE service guidelines
- Sign a DataONE partnership agreement
- Operational implementation by DataONE and Member Node

Response 2: Not at this time

- Link up with appropriate Member Node

  Obtain assistance with data formatting and upload

  Obtain advice about becoming a Member Node in the future.

Step 4:

Following an affirmative response,

- Adhere to communication plan
- Engage in ongoing dialogue with other Member Nodes, Coordinating Nodes and DataONE leadership
- Develop the following documentation:

  Overall MN strategic plan

  5-yr plan for MNs

  MN value proposition

  Roles, responsibilities, value

  Profile registration template

  Service agreement

Rollout Process